BIG4small

Small Area Estimation Conference

Book of Abstracts

Editors: Monica Pratesi, Roberta Siciliano,
Antonio D'Ambrosio, Gaia Bertarelli

Virtual Event, September 20-24, 2021

Contents

Preface

The main purpose of the BIG4small Small Area Estimation Conference 2021 is assess the current state of development and usage of small area methodology.
The aim is to serve as a bridge among statisticians, computer scientists, engineers, and practitioners working on Small Area Estimation in academia, private and government agencies.
This book includes the abstracts of the papers presented at the international conference on Small Area Estimation (SAE) 2021 whose authors paid the registration fee.
All abstracts appear in the book as received. Authors are responsible for the entire content and accuracy of their abstracts.

IS1: Bayesian Statistics for Small Area Estimation

Shrinkage Estimation with Singular Priors and an Application to Small Area Estimation

Malay Ghosh
University of Florida, ghoshm@stat.ufl.edu

The paper considers estimation of the multivariate normal mean under a multivariate normal prior with a singular precision matrix. Conditions for minimaxity of hierarchical and empirical Bayes estimators under such priors are provided. The singular prior is applied to the Fay-Herriot small area estimation model with random effects having the singular distribution. Numerical simulations confirm the suitability of such priors even in a spatial set up.

Small Area Estimation of Vaccination Coverage Using Non-Survey Sources

Trivellore Raghunathan
University of Michigan, teraghu@umich.edu

Small area estimation methods have relied on direct estimates of the population mean (or a proportion) of a continuous (or a binary) variable of interest from one or more surveys. Most methods use hierarchical models to derive the small area estimates. During the last few years, many non-survey data sources have become available that could be used exclusively, or as supplements to surveys, for small area estimation. One example of a potential non-survey source for estimation is Immunization Information Systems (IISs), also known as immunization registries, which are confidential, population-based, computerized databases that record all immunization doses administered by participating providers to persons residing within a given geopolitical area. In this analysis, IIS data from five participating jurisdictions (Michigan, Minnesota, New York City, North Dakota and Oregon) in the United States, augmented by covariates from the National Center for Health Statistics, Health Resources & Services Administration, and the population counts from the U.S. Census Bureau, are used to derive model-based vaccination coverage estimates at the county, state and national levels. Covariates were identified based on a literature review and expert opinion to be predictive of vaccination coverage. A Bayesian hierarchical modeling framework is used in deriving the estimates. We estimated vaccination coverage for a variety of vaccines for children under the age of 18 years. We will present the results for a few vaccinations as examples and compare them with the National Immunization Surveys (NIS) based estimates. Modeling challenges, validation techniques, and other important issues will be addressed. Implications for future research will be discussed.

Pseudo Bayesian Estimation of One-way ANOVA Model in Complex Surveys

Terrance D. Savitsky
U.S. Bureau of Labor Statistics, Savitsky.Terrance@bls.gov

We devise survey-weighted pseudo posterior distribution estimators under two-stage informative sampling of both primary clusters and secondary nested units for a one-way analysis of variance (ANOVA) population generating model as a simple canonical case where population model random effects are defined to be coincident with the primary clusters. We consider estimation on an observed informative sample under both an augmented pseudo likelihood that co-samples the random effects, as well as an integrated likelihood that marginalizes out the random effects from the survey-weighted augmented pseudo likelihood. This paper includes a theoretical exposition that enumerates easily verified conditions for which estimation under the augmented pseudo posterior is guaranteed to be consistent at the true generating parameters. We reveal in simulation that both approaches produce asymptotically unbiased estimation of the generating hyperparameters for the random effects when a key condition on the sum of within cluster weighted residuals is met. We present a comparison with the frequentist EM and a method that requires pairwise sampling weights.

IS2: Time series methods for Small Area Estimation

Multilevel time-series models for estimation at different frequencies and regional levels

Harm Jan Boonstra
Statistics Netherlands, hbta@cbs.nl

A small area estimation method is developed to produce monthly provincial and quarterly municipal unemployment figures. To this end a multilevel time-series model is proposed that uses monthly direct estimates for municipalities and accompanying variance estimates as input. A consistent set of estimates at different aggregation levels is then derived by aggregation of the monthly municipal model-based predictions. The model borrows strength over time and space in several ways. Municipalities belonging to the same province share a common provincial smooth trend, with municipality-specific deviations modeled as local level trends. The model also borrows information from auxiliary series derived from a claimant counts register, with coefficients that can vary over both municipalities and time. The municipal random effects are allowed to vary smoothly over space according to a spatial autoregressive process. To account for the diversity of municipalities and for more volatile time-dependence, non-normally distributed municipal random effects and trend innovations are used via so-called global-local shrinkage priors. A Bayesian approach is taken and the model is estimated by MCMC simulation using R package mcmcsae. It is found that the estimates based on the multilevel time-series model compare favourably to estimates based on a cross-sectional multilevel model.

Multilevel time series modeling of mobility trends in the Netherlands for small domains

Sumonkanti Das
s.das@maastrichtuniversity.nl

The purpose of the Dutch Travel Survey is to produce reliable estimates on mobility of the Dutch population. In this paper, multilevel time-series (MTS) models have been developed to estimate reliable mobility trends of the Dutch population at several aggregation levels, accounting for discontinuities induced by two different redesigns, and outliers due to less reliable outcomes in one particular year. The target mobility variables in this paper are the average number of trip legs per person per day (pppd) and the average distance traveled per trip leg, where trip legs are characterized by motive and transportation modes for a particular journey. The MTS models for the target variables are fitted to annual input series of direct estimates and standard errors at the most detailed breakdown into 504 domains defined by the combination of sex, age-class, motive and mode for the period 1999-2017. Appropriate transformations for the direct estimates and Generalized Variance Functions to smooth the standard errors of the direct estimates are proposed. The models are fitted in an hierarchical Bayesian framework using Markov Chain Monte Carlo (MCMC) simulations by incorporating global-local priors for regularization purposes. Smooth trend estimates at the most detailed domain level are computed from the model outputs and the predictions at higher aggregation levels calculated by aggregation of the most detailed domain predictions result in a numerically consistent set of trend estimates for all target variables. Model diagnostics have also been illustrated for evaluating the fitted models and the corresponding results.

Some Issues in Seasonal Adjustment of Time Series from Repeated Sample Surveys

William R. Bell
U.S. Census Bureau, william.r.bell@census.gov

Time series methods are increasingly being explored and used to improve estimates from repeated sample surveys, a form of small domain estimation. Series from repeated surveys that produce sub-annual estimates, most commonly monthly or quarterly, often exhibit seasonal behavior. Standard practice for statistical agencies publishing such data is to estimate and remove the seasonal pattern from the data so it can be more easily interpreted, a process called seasonal adjustment. With a few rare exceptions, seasonal adjustment practice takes no account of sampling error in the data. The talk will review some issues that arise when seasonally adjusting time series with sampling error. These include (i) the foundational question of what seasonal adjustment is estimating in this case, (ii) how this choice affects variances of seasonally adjusted estimates, (iii) contributions to seasonal adjustment error from the components (seasonal, nonseasonal, and sampling error) and from various estimation error sources (forecast extension error, parameter uncertainty), and (iv) miscellaneous other issues.

IS3: Young Researchers' contribution to Small Area Estimation

Variable and transformation selection for linear mixed models

Yeonjoo Lee; *University of Bamberg, yeonjoo.lee@uni-bamberg.de*
Timo Schmid; *University of Bamberg, timo.schmid@uni-bamberg.de*
Natalia Rojas-Perilla; *Freie University of Berlin, natalia.rojas@fu-berlin.de*
Marina Runge;*Freie University of Berlin, marina.runge@fu-berlin.de*

Variable and transformation selection is a widely studied problem in linear mixed regression models states that the selection of a transformation may be properly viewed as model selection. The working model always depends on which procedure is done first, variable or transformation selection. Although often used for small area estimation, the strategy for selecting the working model under different transformations based on linear mixed models is still under discussion. In this paper we propose a simultaneous variable and transformation selection procedure for linear mixed models based on a conditional Akaike information criterion (cAIC). For this purpose, we adjust the cAIC for linear mixed models proposed in using the Jacobian of the transformation. As the Jacobian adjusted cAIC allows to compare model candidates with differently transformed response variable, all possible model candidates with different sets of independent variables and different transformations can be compared by using the adjusted cAIC. Following this, we introduce a stepwise variable and transformation selection using the adjusted cAIC for increased computational practicability. We include a selection of data-driven transformations into the stepwise selection procedure using adjusted cAIC. Our work involves extensive model-based simulations under different scenarios. Finally, the conclusions from the empirical studies are applied to real data.

Time stable empirical best predictors under a unit-level model

Maria Guadarrama Sanz; *LISER, maria.guadarrama@liser.lu*
Domingo Morales Gonzalez; *Universidad Miguel Hernandez de Elche, d.morales@umh.es*
Isabel Molina Peralta; *Universidad Carlos III de Madrid, isabel.molina@uc3m.es*

Comparability as well as stability over time are highly desirable properties of regularly published statistics, specially when they are related to important issues such as people's living conditions. For instance, poverty statistics displaying drastic changes from one period to the next for the same area have low credibility. In fact, longitudinal surveys that collect information on the same phenomena at several time points are indeed very popular, specially because they allow analyzing changes over time. Data coming from those surveys are likely to present correlation over time, which should be accounted for by the considered statistical procedures, and methods that account for it are expected to yield more stable estimates over time. A unit-level temporal linear mixed model is considered for small area estimation using historical information. The proposed model includes random time effects nested within the usual area effects, following an autoregressive process of order 1, AR(1). Based on the proposed model, empirical best predictors (EBPs) of small area parameters that are comparable for different time points and are expected to be more stable are derived. Explicit expressions are provided for the EBPs of some common poverty indicators. A parametric bootstrap method is also proposed for estimation of the mean square errors under the model. The proposed methods are studied through different simulation experiments, and are illustrated in an application to poverty mapping in Spanish provinces using survey data on living conditions from years 2004-2006.

Controlling the bias for M-quantile estimators for small area

Francesco Schirripa Spagnolo; *University of Pisa; francesco.schirripa@ec.unipi.it*
Gaia Bertarelli; *Sant'Anna School of Advanced Studies, gaia.bertarelli@santannapisa.it*
Ray Chambers; *University of Wollongong, ray@uow.edu.au*
David Haziza; *University of Ottawa, dhaziza@uottawa.ca*
Nicola Salvati; *University of Pisa, nicola.salvati@unipi.it*

Representative outlier units occur frequently in surveys (Chambers, 1986). Several methods have been proposed to mitigate their effects in estimation. If outliers are a concern also for estimation of population quantities, it is essential to pay attention to them in a small area estimation (SAE) context, where some sample sizes are very small and the estimation in often model-based. Chambers and Tzavidis (2006) addressed this issue of outlier robustness in SAE by fitting outlier robust M-quantile models to the survey data. Sinha and Rao (2009) addressed the same issue from the perspective of linear mixed models. Both these approaches use plug-in robust prediction, i.e. they replace parameter estimates in optimal, but outlier-sensitive, predictors by outlier robust versions (a robust-projective approach). These predictors are efficient under the correct model but may be sensitive to the presence of outliers because they use plug-in robust prediction which usually leads to a low prediction variance and a considerable prediction bias. Dongmo Jiongo et al. (2013) and Chambers et al. (2014) proposed a bias correction method for models with continuous response variables. In this talk, we apply two general methods (i.e., for continuous and discrete data) to reduce the prediction bias of the robust M-quantile predictors in SAE context. The first estimator is based on the concept of conditional bias and extends the results of Beaumont et al. (2013) and Favre-Martinoz (2015). Then, we apply a unified approach to M-quantile predictors for continuous and discrete data which is based on a full bias correction. It could be viewed as a generalization of Chambers (1986). An extensive Monte-Carlo simulation study is conducted. Its results confirm that our approaches improve the efficiency and reduce the prediction bias of M-quantile predictors when the population contains units that may be influential if selected in the sample. Data from the EU-SILC 2017 survey in Italy are analysed.

IS4: Issues and opportunities from record linkage and data integration in Small Area Estimation

Record linkage, measurement error and unit level small area estimation: a Bayesian approach

Serena Arima
University of Salento, serena.arima@unisalento.it
Gauri Datta
University of Georgia, gauri@uga.edu

Small area estimation is a statistical techniques involving the estimation of parameters for small sub-populations, generally used when the sub-population of interest is included in a larger survey. Suppose we have m small areas and our goal is to predict the mean of the variable of interest in each area. Using a set of covariates, we consider a well-known unit level model. However, as widely discussed in the literature, covariates might be affected by measurement error and ignoring such an error may lead to misleading conclusions in terms of both parameters estimation and small area mean predictions. In this work, we consider the very common situation in which covariates come from a different data file with respect to the data file of the response variable. As a consequence, covariates are affected by two sources of error: measurement error and linking error. We propose a multivariate nested error regression model that accounts for both sources of error. We conduct a noninformative Bayesian inference by assigning an improper prior that we prove to lead to a proper posterior distribution. The model performance are investigated using different simulation scenarios and with a real data application.

Small area estimation in a linkage errors framework: area-level vs unit-level models

Lordana Di Consiglio
ISTAT, diconsig@istat.it
Tiziana Tuoto
ISTAT, uoto@istat.it

In Official Statistics, interest in data integration has grown enormously in recent years, but the effect of the integration procedures on the statistical analyses has not yet been sufficiently developed. Data integration is not an error-free procedure and linkage errors, such as false links and missed links can invalidate standard estimates. Lately, increasing attention has been addressed to the effect of linkage errors on statistical analyses and statistical predictions. Di Consiglio and Tuoto (2016), Briscolini et al. (2018) propose methods to adjust the unit-level small area estimators for linkage errors when the small area is correctly specified. In this paper, we compare the naive and the adjusted unit-level estimators with the area -level estimators that are unaffected by the linkage errors in the given scenario. The comparison encourages the use of the adjusted unit level estimator even in presence of record linkage errors.

Error in covariates in small area estimation and a generalized Fay-Herriot Model

Gauri Sankar Datta, *University of Georgia and U.S. Census Bureau, gauri@uga.edu*
Kyle M. Irimata, *U.S. Census Bureau,kyle.m.irimata@census.gov*
Jerry Maples, *U.S. Census Bureau,jerry.j.maples@census.gov*
Eric Slud, *U.S. Census Bureau and University of Maryland, evs@math.umd.edu*

The Fay-Herriot model is immensely popular in small area estimation to model area level direct estimates of m unknown small area population means θ_i's by borrowing information from other areas and covariates. It links θ_i's through a multiple linear regression $X_i^T\beta$ on a set of covariates X_i. In addition to the linearity of the regression function, the model also assumes normality and homoscedasticity of the model error components. In various applications some of these assumptions fail to hold. Specific examples include non-normality of the model errors or a non-linear link function, $\mu_i(\beta)$, of $X_i^T\beta$. Also measurement errors in covariates transform a regular Fay-Herriot model to a model with heteroscedastic model errors. In this work, we generalize the standard Fay-Herriot model by allowing the link function non-linear, or model error non-normal or heteroscedastic. We compute the empirical best (EB) prediction of θ_i based on a convex combination of Y_i and the estimated mean function $\hat{\mu}_i$. We estimate the model parameters using estimating equations depending linearly on Y_i and $Y_i^2, i = 1, \cdots, m$. Under moments assumption on the model error terms we investigate the asymptotic (in m) mean squared error (MSE) of the EB predictors. Our results include some known results as special cases under additional distributional assumptions for the model errors. We adapt an existing nonparametric bootstrap method to accurately estimate the MSE. We conduct an extensive simulation study under non-normality of the Fay-Herriot model error terms and compare the proposed bootstrap estimator with two popular MSE estimators due to Lahiri and Rao (1995) (see also Prasad and Rao, 1990) and Datta et al. (2005).

IS5: Selected Challenges is Small Area Estimation

Empirical best prediction of bivariate nonlinear small area indicators

Domingo Morales, *Universidad Miguel Hernandez de Elche, d.morales@umh.es*
Maria Dolores Esteban, *Universidad Miguel Hernandez de Elche, md.esteban@umh.es*
Maria José Lombardia, *Universidade da Coruna, maria.jose.lombardia@udc.es*
Esther Lopez Vizcaino, *Instituto Galego de Estatistica, esther.lopez@ige.eu*
Augustin Pérez, *Universidad Miguel Hernandez de Elche, agustin.perez@umh.es*

This work introduces a bivariate nested error regression model for estimating bivariate nonlinear small area indicators, with special emphasis on ratio indicators defined as sums of ratios or as ratios of sums. Based on the new model, empirical best predictors are introduced and parametric bootstrap procedures for estimating mean squared errors are proposed. Several simulation experiments, designed to analyse the behaviour of the introduced fitting algorithm, predictors and mean squared error estimators, are carried out. An application to real data from the Spanish Household Budget Survey illustrates the behaviour of the proposed statistical methodology. The target is the estimation of ratios of food household expenditures by Spanish provinces.

On "qape" R package for measuring accuracy of small area predictors

Thomasz Zadlo
University of Economics in Katowice, tomasz.zadlo@uekat.pl
Alicja Wolny-Dominiak
University of Economics in Katowice, woali@ue.katowice.pl

In the package the problem of prediction based on linear mixed models (LMM) is implemented. It can be used both for cross-sectional and longitudinal data. The following three predictors are taken into account. Firstly, we have included the empirical best linear unbiased predictor of a linear combination of the variable of interest for models covered by lmer R-function (including models with correlated random effects and with heteroscedastic random components). Secondly, we have implemented plug-in predictors of any given population or subpopulation characteristic for the same class of models additionally including any given transformation of the variable of interest. Thirdly, the empirical best predictor of any given population or subpopulation characteristic under nested error linear mixed model is taken into account. The prediction accuracy of the predictors, measured by MSE and QAPE (quantile of absolute prediction error), is estimated using parametric bootstrap, residual bootstrap and double bootstrap methods. New proposals of MSE and QAPE estimators based on the double bootstrap method are presented.

Regularized Small Area Estimation: A Framework for Robust Estimates in the Presence of Unknown Covariate Measurement Errors

Joscha Krause
Trier University, krause@uni-trier.de

SAE provides stable estimates of area statistics in the presence of small samples. This is achieved by combining observations from multiple areas in suitable regression models. These models exploit the functional relation between the area statistic and contextually related covariate data to make predictions for the quantities of interest. An important assumption of this methodology is that the covariate data is measured correctly. If this does not hold, area statistic estimates can be severely biased or highly inefficient. In that case, methodological adjustments are required to allow for reliable results. There are several approaches in the literature that allow for robust estimates from contaminated data bases. Unfortunately, many of them share a common limitation. Robust SAE techniques typically require distribution assumptions on the measurement error. These assumptions can be either explicit by requiring a specific distribution, or implicit by demanding that the distribution is known. However, both settings are rarely verifiable in practice. We propose a new approach to robust SAE that does not require distribution assumptions on the measurement error. Using insights into robust optimization theory, we proof that regularized model parameter estimation is equivalent to the robust minimization of loss functions under arbitrary model matrix perturbations. This equivalence holds for many well-established regularized regression methods, such as the LASSO, ridge regression, and the elastic net. It allows us to produce reliable area statistic estimates in the presence of unknown covariate measurement errors. We built upon this result to derive a modified Jackknife algorithm that allows for conservative MSE estimation for predictions obtained on contaminated data bases. In addition to that, we discuss consistency in model parameter estimation of regularized regression in this setting. The effectiveness of the methodology is demonstrated in a Monte Carlo simulation study.

IS6: Disaggregated data and indicators from Big data sources

Social networks data and small area estimation: a tentative solution to overcome selection bias

Elena Siletti, *University of Milano, elena.siletti@unimi.it*
Stefano Maria Iacus, *University of Milano, stefano.iacus@unimi.it*
Giuseppe Porro, *University of Insubria, giuseppe.porro@uninsubria.it*
Silvia Salini, *University of Milano, silvia.salini@unimi.it*

Social networks are a source of large and continuous flows of information, opinions and debates that produce a kind of the so called big data. That is the reason why social media has been recently considered the largest focus group in the world. The opinions expressed on social networks are continuously updated, cover all kind of topics, involve people from different social strata, are (mostly) free from censorship, data form social media are more timely and possess higher space granularity, and, last but not least, they come to social analysts in huge amount for free or inexpensive. Despite these and others important positive traits, this data revolution bring with it some problems, especially one of the major criticisms refers to the selection bias. Since it is clearly not possible to renounce to the use of such a rich source of data, and the right way can only be to face new methodological challenges, in this paper, we suggest to mash-up official statistics with social media data, with the purpose of measuring and taking into account this bias. In detail, we propose to adjust statistics based on Twitter data by anchoring them to reliable official statistics through a weighted, space-time, small area estimation model. This method can be used anytime official statistics exist at proper level of granularity and for which social media usage within the population is known. An empirical application, based on the use of an innovative measure for subjective well-being at work and derived from the analysis of Italian Twitter data from 2012 to 2017, is performed as an example. The weights depend on broadband coverage and Twitter rate at province level, while some demographic and economic variables at regional level are adopted in the space-time small area estimation model. The resulting statistics are then compared with those by survey on the quality of job at macro-economic regional level provided by Italian Office of Statistics (ISTAT), showing evidence of similar paths.

Small area poverty indicators adjusted using local price indexes.

Luigi Biggeri, *University of Florence, luigi.biggeri@unifi.it*
Caterina Giusti, *University of Pisa, caterina.giusti@unipi.it*
Stefano Marchetti, *University of Pisa, stefano.marchetti@unipi.it*
Monica Pratesi, *University of Pisa, monica.pratesi@unipi.it*

We focus on estimating monetary poverty indicators at sub-regional level in Italy taking into account the different price levels within the country. To account for the local price levels, Spatial Price In- dexes (SPIs) are computed using retail scanner data on regional and sub- regional retail volumes (units) and price for food and grocery. Specifically, a Country Product Dummy model is used, with products aggregated by province and ECOICOP-8-digit classification, for a total of 103 provinces and 102 ECOICOP-8-digit. The SPIs are used to adjust the poverty line when computing provincial poverty indicators that are estimated using area level Small Area Estimation (SAE) models, which link direct unreli- able estimates to aggregated auxiliary information, often easily available. The use of SAE models is necessary to obtain reliable estimates at sub- regional level that can be used to guide policy decisions to reduce poverty.

Small area estimation via Heteroskedastic Geographically Weighted Regression for functional data

Elvira Romano, *University of Campania "Luigi Vanvitelli", elvira.romano@unicampania.it*
Diana Andrea, *University of Campania "Luigi Vanvitelli", andrea.diana@unicampania.it*
Jorge Mateu, *University Jaume I, mateu@mat.uji.es*

Small area estimation is studied under a Heteroskedastic Geographically Weigthed Regression model for functional data . The spatio-functional model we introduce assumes that the variance varies across the space, and that each local model (defined at each location) gives a local non parametric estimation of the variance. This approach improves the model performance in terms of predictive fit for small area estimation, as illustrated by simulations and through the analysis of a real data set.

IS7: Small area estimation for latent variables and complex indicators

Estimating small area latent social integration of second-generation students in Italy

Francesco Giovinazzi
University of Bologna, francesco.giovinazz2@unibo.it
Daniela Cocchi
University of Bologna, daniela.cocchi@unibo.it

The existence of cultural divides and prejudices complicate the processes of integration and acculturation of migrant families, and it may affect in particular some nationalities over the others. The education system has a leading role in fostering and attaining social integration, in particular when it comes to the younger sections of the migrant population, otherwise known as second generation. We propose to use data from the survey on Integration of the second generations (ISG), carried out by the Italian National Institute of Statistics (Istat) in 2015, to assess the level of foreign students social integration according to nationality and geographic area. Here we propose a 2-step approach. In the first step, we select the most suitable items representing the different dimensions of social integration and we use a latent class model to cluster second-generation children into homogeneous groups in terms of multivariate latent social integration. We then aggregate the information according to unplanned domains, obtained by combining the nationality of the students with the administrative region in which they attend the school. We can interpret the proportion of students belonging to the most integrated class, in each domain, as a multivariate indicator of latent social integration. In the second step, we use a small area model to improve the estimate of the proportion in the domains, borrowing strength from administrative data coming from the Official Register of the Ministry of Education. The result is a synthetic indicator taking into account the many different dimensions of social integration, such as education, relationships with friends and family, language, and household conditions.

Unit level models on the log-scale: a new Bayesian proposal for poverty mapping

Enrico Fabrizi, *Università Cattolica del Sacro Cuore, enrico.fabrizi@unicatt.it*
Aldo Gardini, *University of Bologna, aldo.gardini2@unibo.it*
Carlo trivisano, *University of Bologna, carlo.trivisano@unibo.it*

Mixed models are popular tools for small area estimation when a model linking the target variable and linking information can be specified at the unit level. In many applications positive variables with positively skewed distributions are often of interest. For instance, in poverty and inequality mapping, target parameters are summaries of the distribution of income, consumption or other size variables. A popular approach in this case is to specify a unit level model for the variable in question and then use predicted values for out-of-sample units to estimate poverty mapping measures at the small area level. Unfortunately, normality of the residuals and other random components is often not tenable. Among various solutions to sidestep this problem, to specify a normal linear mixed model on a normalizing transformation of the target variable is considered in this research; namely we focus on the log transformation in line with many other papers in the literature. When following a Bayesian approach, the analysis of linear models on the log-scale deserve special attention as the specification of standard priors of the variance of the residuals (and other variance components) lead to predictive distributions that although proper have no finite moments, thus preventing the application of ordinary summary measures such as the mean. Following the findings of the same authors in previous papers, we characterize the finiteness of the posterior predictive distribution moments when a normal linear mixed model is specified on the log scale. Specifically, we focus on generalized inverse Gaussian priors for the variance components and derive conditions on the prior parameters that guarantee finite moments for the predictive distribution up to a pre-specified order. Moreover, we propose a strategy for the choice hyper-parameters that lead to a uniform prior for the intraclass correlation coefficient, and more in general allows to specify a prior balance between the variance components. The frequentist properties of small area predictors associated to the discussed model specification are evaluated in a simulation study where they are compared to models adopting priors that do not guarantee the existence of posterior moments for the predictive distribution and to the current proposals in the empirical Bayes (EB) literature on the topic. According to our results, the suggested prior specification produces estimates which outperform the other hierarchical Bayes

solutions and are close to the performances of EB procedure in terms of bias, MSE, but also frequentist coverage and average width of the posterior intervals. Eventually, an application to a real scale data set is considered with the purpose of illustrating computational and practical computational issues.

Multivariate small area estimation methods for multidimensional latent wellbeing indicators

Angelo Moretti
Manchester Metropolitan University, a.moretti@mmu.ac.uk
Natalie Shlomo
University of Manchester, natalie.shlomo@manchester.ac.uk

Multivariate small area estimation is an ongoing research field. In the study of poverty and wellbeing phenomena, policy makers ask for reliable information about the geographical distribution of social indicators. Since these phenomena are multidimensional, a set of indicators needs to be studied. Since these are likely to be correlated, the problem of multivariate statistical distributions arises. In this work, we study the estimation of multidimensional social exclusion indicators in a multivariate small area estimation setting under the multivariate Empirical Best Linear Unbiased Predictor (EBLUP) approach based on the Harter and Fuller multivariate mixed-effect model. In particular, a dashboard of predictions and composite measures estimated from factor analysis models are investigated and compared via simulation and real data applications. Furthermore, in this paper, we will also discuss the problem of mean squared error estimation in presence of latent variables in small area estimation. In fact, the variability arising from the use of factor analysis models, or other data dimensionality reduction methods must be taken into account when model-based mean squared error estimates are produced, otherwise we may under estimate them. This paper is an extension of our previous work where we studied the use of univariate EBLUP in latent economic wellbeing indicators measurement, showing that the use of latent factor scores provides multidimensional estimates with a lower variability than simple and weighted averages of a dashboard of indicators. Applications using data from the European Union Statistics on Income and Living Conditions (EUSILC) are presented. These are related to economic wellbeing, and more specifically to housing deprivation in Italy. With this work we aim to provide guidance to policy makers and local governments in charge of wellbeing and poverty decisions who need reliable estimates at a local level. We stress the problem of multivariate small area estimation due to correlation in wellbeing and poverty data. This might yield more precise estimates than univariate small area estimation.

IS8: Small Area Estimation in Official Statistics

Small area estimates of labour market status using multinomial expectile regression

James Dawber
University of Southampton, J.P.Dawber@soton.ac.uk
Enrico Fabrizi
Università Cattolica del Sacro Cuore, enrico.fabrizi@unicatt.it

M-quantile approaches for small area estimation have been shown to be a useful alternative to mixed models when certain assumptions cannot be made. We present an expectile modelling approach for when the response is a categorical outcome, such as the three categories of the labour market status. This improves on common SAE methods that only estimate the proportions of people in subregions that are unemployed. We show how this novel approach is much faster and reliable compared to multinomial mixed model approaches.

Robust small area estimation in business surveys

Chiara Bocci, *University of Florence, chiara.bocci@unifi.it*
Paul A. Smith, *University of Southampton, P.A.Smith@soton.ac.uk*
Nikos Tzavidis, *University of Southampton, N.TZAVIDIS@soton.ac.uk*
Sabine Krieg, *Statistics Netherlands, s.krieg@cbs.nl*
Marc J.E. Smeets, *Statistics Netherlands, mje.smeets@cbs.nl*

Small area estimation is still rarely applied in business surveys, because of challenges arising from the skewness and variability of many size-related variables. Sampling from skewed variables produces samples containing outliers, which affect the models on which small area estimates are based. A number of different approaches have been introduced to robustify these models to improve the properties of small area estimates. Here we apply a range of small area estimation methods to real business data in a situation where the population is known: we use Netherlands firms tax register data and a sampling procedure which replicates the sampling for the retail sector of Statistics Netherlands Structural Business Survey as a basis for investigating the repeated sampling properties of small area estimators. In particular, we consider the use of the EBLUP under a random effects model and variations of the EBLUP derived under (a) a random effects model that includes a complex specification for the level 1 variance and (b) a random effects model that is fitted by using the survey weights. Although accounting for the survey weights in estimation is important, the impact of influential data points remains the main challenge in this case. Then we explore the use of outlier robust estimators in business surveys, in particular (a) a robust version of the EBLUP, (b) M-regression based synthetic estimators, and (c) M-quantile small area estimators. The latter family of small area estimators includes robust projective (without and with survey weights) and robust predictive versions. M-quantile methods have the lowest empirical mean squared error and are substantially better than direct estimators, though there is an open question about how to choose the tuning constant for bias adjustment in practice. Finally, we explore a doubly robust approach comprising the use of survey weights in conjunction with outlier robust methods in small area estimation.

Causal inferences for official statistics

Setareh Ranjbar, *University of Lausanne HEC, setareh.ranjbar@unil.ch*
Nicola Salvati, *University of Pisa, nicola.salvati@unipi.it*
Barbara Pacini, *University of Pisa, barbara.pacini@unipi.it*

Small area estimation provides official statistics that are vastly used by policy makers. These analyses even though are not mainly based on proper causal assumptions, they intend to be interpreted in that way. On the other hand, when doing impact evaluation in many cases, it is important to acknowledge the heterogeneity of the treatment effects for different domains. Where certain geographic, socio-demographic, or socio-economic domains may benefit from a program/ policy intervention, others may be worse off. Such analysis can help the governments to put in place localized rather than global interventions that can more efficiently target the help seekers in the society. In this setting if the domain for which we are interested in the impact, is small with regards to its sample size (or even zero in some cases), then the evaluator has entered the small area estimation dilemma. Based on the modification of the Inverse Propensity Weighting estimator and the traditional small area predictors, we propose a new methodology to estimate area specific average treatment effects for unplanned domains. These methods enable us to estimate the impact even for those small domains where all units in the sample belong to treated or control group, provided that there information for both groups in the auxiliary variables at the population level. By means of these methods we can also provide a map of policy impacts, that can help to better target the treatment group(s). Further, we develop analytical Mean Squared Error (MSE) estimators of the proposed predictors. The results of our simulations show a clear gain of the proposed techniques to the existing methods. In the end we provide a real life example where we estimate the effects of permanent versus temporary contracts on the economic insecurity of households in different regions of Italy.

IS9: Recent Advances in Model Selection and Diagnostics for Small Area Estimation

Selection of auxiliary variables for two-fold subarea-level linking models in small area estimation: A simple method

J. N. K. Rao
Carleton University, jrao34@rogers.com
Song Cai
Carleton University, scai@math.carleton.ca

Model-based small area estimation, using area-level data, depends on a sampling model on the direct estimators of area means and a model linking the area means to area-level auxiliary variables. When multiple auxiliary variables are available, it is desirable to use a variable selection method to select a parsimonious linking model that fits the data well. For the basic Fay-Herriot (FH) area level model, Lahiri and Suntornchost (2015) proposed a simple method of variable selection by estimating a standard information criterion under the assumed linking model. In this paper, we extend their work to two-fold subarea-level linking models using a parameter-free transformation method to estimate the information criteria. We report the results of a simulation study on the performance of the proposed method relative to other methods that ignore the area random effect and use only subarea-level random effect in the specification of the linking model. The main advantage of two-fold linking models is that the resulting subarea estimators for non-sampled subareas take advantage of the area level random effect in the model and lead to significant gains in efficiency over the synthetic estimators based on the one-fold model based only on the subarea random effect.

A Robust Goodness-of-fit Test for Small Area Estimation

Mahmoud Torabi
University of Manitoba, Mahmoud.Torabi@umanitoba.ca
Jiming Jiang
University of California, jiang@wald.ucdavis.edu

We develop a method originally proposed by R. A. Fisher into a general procedure, called tailoring, for deriving goodness-of-fit tests that are guaranteed to have a chi-square asymptotic null distribution. Furthermore, the method has a robustness feature that it works correctly in testing a certain aspect of the model while some other aspects of the model may be misspecified. We apply the method to small area estimation (SAE) for detecting potential model misspecification. A test is proposed using the tailoring method that incorporates the special interest of SAE. We evaluate the performance of the tests both theoretically and empirically and compare the performance with several existing methods. Our empirical results suggest that the proposed test is more accurate in size, and has either higher or similar power compared to the existing tests. The proposed test is also computationally more effective than the sensitivity test that is among the comparing methods. A real data application is considered

Recent Advances in Measures of Uncertainty in Post Model Selection Small Area Estimation

Thuan Nguyen, *Oregon Health & Science University, nguythua@ohsu.edu*
Mahmoud Torabi, *University of Manitoba, Mahmoud.Torabi@umanitoba.ca*
Jiming Jiang, *University of California, jiang@wald.ucdavis.edu*

Small area estimation (SAE) utilizes statistical models in order to borrow strength. Therefore, model selection is often a standard practice in SAE to make sure that an appropriate model is used before producing small area estimates based on the model that is chosen. A measure of uncertainty must take into account the (additional) uncertainty in model selection. In this presentation, I will discuss some recent advances in obtaining measures of uncertainty in post model selection SAE. These include a Monte-Carlo jackknife (McJack) method and a recently proposed Sumca method. while McJack produces a second-order unbiased estimator for the logarithm of the mean squared prediction error (MSPE) of a small area predictor, Sumca leads to a second-order unbiased estimator of the MSPE, and has computational advantage over the McJack.

IS10: Small Area Estimation for Permanent population Census and Social Surveys:new applications and methods

MIND, an R package for multivariate small area estimation with multiple random effects

Stefano Falorsi, *ISTAT, stfalors@istat.it*
Michele D'Alò, *ISTAT, dalo@istat.it*
Andrea Fasulo, *ISTAT, fasulo@istat.it*

The proposed Small Area Estimator is based on a multivariate and multi random effects linear mixed model implemented through the R function MIND (Multivariate model based INference for Domains), developed by Istat. The presented method may be seen as an extended multivariate version of the more standard linear mixed model at unit level. The main extensions give the possibility to consider two or more random effects in the model and to consider a multivariate qualitative variable as dependent variable, following the multivariate modelling approach of Datta, Day e Basawa (1999). The introduction of the marginal random effects is very useful when many areas are not sampled, in as much as it limits the bias of the synthetic estimator. More specifically, when some domains of interest are not observed in the sample (as they are not planned) or are strongly under represented, it is possible to introduce one or more marginal random effects that, on the contrary, are all observed in the sample. The model can be expressed as follows $y = X\hat{I}^2 + Zu + e$.

The estimates obtained with MIND fall within the group of estimators classified as model-based General Projection (GP). The general formulation of GP estimator of the vector of totals for domain d is given by the sum of predicted vector values, on the basis of the above mentioned model, for all the units of the target population falling within the d domain. MIND considers too a particular type of Projection, CP, estimator where predicted values are used only for the r subgroup of units that are not included in the sample. The estimate of variance components of the mixed effects model is calculated by a multivariate extension of the REML procedure proposed by Saei e Chambers (2003). In order to produce the labour market indicators at City and Fua level, the potential of the multivariate approach proposed by MIND has been exploited in two ways: to specify the dependent variable y and to define the random effects.

SAE estimation under coherence for different overlapping areas. An application for the estimation of employment and unemployment from LFS for cities and FUAs

Silvia Loriga, *ISTAT, siloriga@istat.it*
Daniela Filipponi, *ISTAT, dafilipp@istat.it*
Michele D'Alò, *ISTAT, dalo@istat.it*

The aim of this job is to describe the statistical methodology used to produce estimates of a selection of labour market variables at City and Fua level and to analyse the results obtained. The parameters, based on the LFS and referring to year 2018, are the following: number of unemployed persons, total and by sex, number of economically active persons, total and by sex, number of economically active persons aged 20 to 64, total and by sex, number of employed persons aged 20 to 64, total and by sex. The estimates of these indicators are calculated through a unit level multivariate model, designed to allow the estimation of the variables of interest in a coherent way. Area univariate models have also been tested, however the use of a specific model for each variable does not guarantee straightly the numeric coherence among estimates. The estimator used is based on a multivariate model implemented through the R MIND function, developed by Istat. The method described in the present work may be seen as an extended multivariate version of the more standard linear mixed model at unit level. The main extensions give the possibility to consider two or more random effects in the model and to consider a multivariate qualitative variable as dependent variable, following the multivariate modelling approach of Datta, Day e Basawa (1999). Specifically, with regards to the random effects, when some domains of interest are not observed in the sample (as they are not planned), it is possible to introduce one or more marginal random effects corresponding to areas or classification variables never out of sample. The dataset created for estimating the indicators is based on Istat Base Register of Individuals and on the Thematic Labour Registry, used as main sources of information for indviduals anagraphich characteristics and working condition. In addition, the dataset has been integrated with information on social aspects, welfare benefits and type of income deriving from welfare agencies, Italian Ministry of Finance and the national revenue agency. This data was essential for introducing auxiliary variables in the models. In order to produce the labour market indicators at City and Fua level, the potential of the multivariate approach proposed by MIND has been exploited in two ways: to specify the dependent variable and to define the random effects. The dependent variable was defined as a vector composed by three dicotomous variables representing the categories

of the employed, unemployed and inactive individuals. The coherence of indicators across different domains was reached via a single cross-classification model that included all the domains of interest. Finally, two models have been defined: one for the Fua estimates and another for the City estimates. The results show efficiency gains with respect to the direct estimates, with particular regards to the estimation of the unemployed persons for which the sample errors were rather high.

Defining the sample designs for small area estimation

Piero Demetrio Falorsi, *Sapienza University of Rome, piero.falorsi@gmail.com*
Stefano Falorsi, *ISTAT, stfalors@istat.it*
Paolo Righi, *ISTAT, parighi@istat.it*

The small area problem is usually considered to be treated via estimation. However, if the domain indicator variables are available for each unit in the population, there are opportunities to be exploited at the survey design stage. This condition is usually met in the business survey context, where the domain indicator variables are available in the business register. The circumstance is respected even in the households surveys for the geographical domains. Singh, Gambino and Mantel (1994) noted a need to develop an overall strategy that deals with small area problems, involving both planning sample design and estimation aspects. In this framework, it is crucial to control the sample size for each domain of interest so that it is treated as a planned domain at the design stage. It is possible to produce direct estimates with a prefixed level of precision. In general, with a design-based approach to the inference, the presence of sample units in each domain allows one to compute domain estimates, although not always reliably. Furthermore, in the model-based or model-assisted approach, sample units in each estimation domain allow one to use models with specific small area effects, giving more accurate estimates of the parameters of interest at the small area level (Lehtonen, Sarndal and Veijanen 2003). Indeed, having sampling units in each domain of interest would also benefit the computation of indirect estimates by enabling a substantial reduction of model bias. Traditional sampling techniques address data disaggregation by oversampling or introducing a more profound stratification. More sophisticated techniques allow improving sampling designs by geographically spreading the sample units (Grafstorm, Lundstram and Schelin, 2012) and diminishing the level of clustering. These approaches would foster reaching segregated or rare subpopulations. In this paper, we consider the problem of estimating the totals Y_d of a variable y for various overlapping domains. With reference to the domain d (being $d = 1, \ldots, D$), we consider a general small-area model $y_{id} = x_{id}\hat{I}^2 + u_d + e_{id}$, where regarding the unit i in the domain d, y_{id} indicates the value of the target variable y, x_{id} denotes a vector of auxiliary variables, $\hat{I}^2$ is a vector of unknown super-population parameters, u_d indicates a random domain effect, and e_{id} a random noise. We focus on the definition of the minimum cost sample design ensuring that the model variances are lower than pre-fixed thresholds. The proposed sampling algorithms is a modification of that illustrated in Falorsi and Righi (2015).

IS11: Small Area estimation: new developments and applications

A Hierarchical Bayesian Approach for Addressing Multiple Objectives in Poverty Research for Small Areas

Stefano Marchetti, *University of Pisa, stefano.marchetti@unipi.it*
Monica Pratesi, *University of Pisa, monica.pratesi@unipi.it*
Gaia Bertarelli, *Sant'Anna School of Advanced Studies, gaia.bertarelli@santannapisa.it*
Partha Lahiri, *University of Maryland, plahiri@umd.edu*

Nowadays the information extracted from data should be the key to good policy and to good decisions, therefore, analysts must make the best possible use of all available information. However, data availability often is limited by cost, by sensitivity of the questions, or for other reasons. As a consequence, there is the need to use data from different sources. Our goals are to develop appropriate large parametric models and hierarchical models and to demonstrate their ability to improve inferences about quantities for which there are meager data. When a hierarchical model can be found to represent the situation properly, analysis of that model often can be used to extract most or all of the relevant information and so provide the best possible estimates. The application considered will include small area estimation in the context of the European Union Statistics on income and living conditions. In developing the hierarchical model, we use together survey data and population registers. Different levels of the model and specification of the covariates used at different levels will be justified with statistical tests of fit appropriate for multi-level models. As for the implementation of the hierarchical model, we propose to use Bayesian methodology assisted by Monte Carlo Markov Chain.

Best Prediction of Missing Area-Level Direct Estimates via Multivariate Modelling

Anna Lena Wolwer, *Trier University, woelwer@uni-trier.de*
Jan Pablo Burgard, *Trier University, burgardj@uni-trier.de*
Domingo Morales, *University Miguel Hernandez de Elche, d.morales@umh.es*
Ralf Muennich, *Trier University, muennich@uni-trier.de*

Model-based small area predictors are generally derived for complete data files. In application to real data, however, data files often contain missing values. In the area-level setting, missing direct estimates of interest are usually predicted with synthetic estimators. Not only is the synthetic estimator not an empirical best predictor, estimating the associated mse is also cumbersome. We extend the multivariate Fay-Herriot model by allowing for arbitrary missing direct estimates of interest and give fitting algorithms for the model parameters. Based on the extended model, we introduce an empirical best predictor of the missing area-level direct estimates. Furthermore, we derive an analytical approximation to its mean squared error. The proposed estimator handles both correlated and uncorrelated sampling errors of direct estimates as well as arbitrary missing structures. It therefore lends itself to a broad range of practical applications. In a simulation study, we undermine the theoretical findings of the proposed estimator. Special emphasis is put on the performance of the best predictor of the missing direct estimates and their mean squared error.

Small area estimation via multivariate generalized linear mixed effects models

Emilia Rocco
University of Florence, emilia.rocco@unifi.it
Maria Francesca marino
University of Florence, mariafrancesca.marino@unifi.it

The analysis of complex phenomena often requires the estimation of correlated descriptive measures, which may potentially have a heterogeneous nature: binary variables, counts, continuous skewed variables, or a combination of them. Moreover, a frequent issue is that of deriving estimates for small spatial domains that are non-sampled or are under-sampled in surveys. Multivariate, unit-level, SAE models have not been studied much. In this work, we suggest the use of a multivariate mixed effect model, based on correlated random effects, for the jointly modelling of multiple outcomes recorded on a sample of units clustered within domains. This allows us to account for the multivariate dependence among outcomes by means of the latent terms in the model. The proposed approach allows for a deeper understanding of the phenomenon under investigation thanks to the estimated correlation between the latent effects; this, in turn, provides an indirect measure of the dependence between the outcomes themselves. From the other side, parameter estimates are improved by borrowing strength across spatial regions and multiple outcomes, simultaneously. Predictions of small area parameters are derived and a parametric bootstrap method is proposed for estimating the mean squared error of such predictions. The proposal is tested by means of an intensive simulation study considering different types of outcomes.

IS12: Some novel developments in small area estimation

Robust, high-dimensional data linkage for small area statistics

Snigdhansu Chatterjee
University of Minnesota, chatt019@umn.edu

The auxiliary information that is often found to be useful for small area modeling may be contained in different datasets. This situation may arise due to ownership and data collection schemes, privacy and confidentiality considerations, or data security requirements. Also, in many modern instances, the data used for small area modeling is high-dimensional in nature. Additionally, such big data may contain outliers and aberrant observations, consequently most statistical techniques need to be used with caution when analyzing such big data. In this work, we present a robust and computationally simple technique for linking high-dimensional datasets. We present theoretical foundations for our proposed methodology, and illustrate using numeric examples.

Covariance based Moment Equations for Improved Variance Component Estimation

Sanjay Chaudhuri
National University of Singapore, stasc@nus.edu.sg

ANOVA-type estimators of variance components for nested error regression models are always constructed based on moment equations related to residual variance. We consider moment equations associated with covariance and construct improved ANOVA-type estimators. These estimators are seen to be consistent, asymptotically unbiased and have better performances than traditional estimators of variance components for almost all kinds of sample allocations. Their improved performance is demonstrated analytically as well as through detailed simulation studies and applications to real data sets.

IS13: Data Science Methodology Transfer: Big to Small

Bias versus statistical errors in Big data information systems

Edwin A. Valentijn

Using proper scoring rules to derive well calibrate photometric redshift models

Kai Polsterer

Error Mitigation in Quantum Measurement through Fuzzy C-Means Clustering

Autilia Vitiello
University of Naples Federico II, autilia.vitiello@unina.it

Recently, Quantum Computing is entered in the so-called Noisy Intermediate-Scale Quantum (NISQ) era, where devices characterized by a few number of qubits are potentially able to overcome classical computers in performing specific tasks. However, noise in quantum operators still limits the size of quantum circuits that can be run in a reliable way. Consequently, there is a strong need for error mitigation approaches aimed at increasing reliability in quantum computation and making this paradigm really useful and productive in real world applications. In this paper, a fuzzy method, such as Fuzzy C-Means (FCM) clustering, has been used, for the very first time, to support the identification of matrices for error mitigation in quantum measurement. As shown in experiments, mitigation matrices identified with the support of FCM are able to strongly reduce error in computation when compared to mitigation matrices conventionally identified, like those used by IBM in its quantum library named Qiskit.

IS14: Latent variables in small are models: theoretical and applied issues

Empirical Best Prediction for Small Area Estimation of categorical variables using Finite Mixtures of Multinomial Logistic Models

Maria Giovanna Ranalli, *University of Perugia, giovanna.ranalli@unipg.it*
Maria Francesca marino, *University of Florence, mariafrancesca.marino@unifi.it*
Nicola Salvati, *University of Pisa, nicola.salvati@unipi.it*
Marco Alfò, *Sapienza University of Rome, marco.alfo@uniroma1.it*

Many survey variables are categorical in nature and SAE methods based on linear mixed models may not be fully appropriate. Jiang (2003) developed an empirical best prediction (EBP) method for general responses in the Exponential Family. In this approach, the area-specific random effects are assumed to be independent and identically distributed (i.i.d.) draws from a Gaussian distribution. One of the drawbacks associated with this assumption entails the computational burden required to derive parameter estimates, compute the EBP and, in particular, provide the corresponding measure of reliability. For non-Gaussian responses, we need to deal with (possibly) multiple integrals that do not admit a closed form expression and, therefore, need to be approximated. To avoid computational issues, ad hoc alternatives, mainly based on plug-in predictors and Taylor linearizations, were proposed and are currently largely applied [Gonzalez-Manteiga et al. (2007), Molina, Saei and Lombardia (2007), Saei and Chambers (2003), Lopez- Vizcaino, Lombardia and Morales (2013)]. In particular, Molina, Saei and Lombardia (2007) take a plug-in approach to estimate small area indicators of labor force participation under a multinomial logit mixed model. In this talk, we propose to leave the distribution of the area-specific random effects (the mixing distribution) of this model unspecified and estimate it from the observed data via a nonparametric maximum likelihood approach [NPML, Laird (1978), Simar (1976), Lindsay (1983a, 1983b)]. This estimate is known to be a discrete distribution defined over a finite number of locations leading to a finite mixture model. The proposed approach is an extension to multinomial models of the approach proposed by Marino et al. (2019) for generalized linear models with general responses in the Exponential Family that offers a number of advantages. First, it allows us to obtain an EBP and not a plug-in approximation and to avoid unverifiable assumptions on the random effect distribution; second, since mixture parameters are directly estimated from the data and are completely free to vary over the corresponding support, extreme and/or asymmetric departures from the homogeneous model can be easily accommodated. Last and more important, the discrete nature of the

mixing distribution allows us to avoid integral approximations and considerably reduces the computational effort.

Small area models with uncertainty on measurement error in covariates

Silvia Polettini
Sapienza University of Rome, s.polettini@gmail.com
Serena Arima
Università del Salento, serena.arima@unisalento.it

In model based small area estimation the borrowing of strength is achieved through the definition of mixed effects regression models that link the small areas. Under this approach, auxiliary information, in the form of model's covariates, play a key role in small area estimation. In the presence of good covariates, it is expected that random small area effects will have small variation, leading to significant shrinkage to the synthetic regression estimator. However, when explanatory variables are measured with error, the standard small area model is not appropriate. Indeed, it has been proven that parameter estimates may be dramatically biased and the variability of the random effects and, consequently, of the small area means significantly increases. The effect of measurement error in covariates may be corrected by specifying a suitable measurement error model. However, sometimes the presence of measurement error in a variable is uncertain; moreover, accounting for measurement error may introduce unnecessary variability when the measurement error is small compared to the sampling error. In this contribution we focus on a Bayesian nested error linear regression model under measurement error in auxiliary variables and introduce uncertainty about the presence of the measurement error using suitable priors, investigating a model-based variable selection approach to decide whether allowing for measurement error or resorting to a standard small area model.

Bayesian model selection for log-linear latent class models

Davide Di Cecco
Sapienza University, davide.dicecco@uniroma1.it

Very often, data disseminated in official statistics consists of unit counts in certain domains, that is, contingency tables reporting the number of units in a population of interest in certain domains. Census hypercubes constitute a notable example. The dimensions of the hypercube represent the categorical variables defining the spreading domain. Whenever the sample size is not large enough to guarantee reliable direct estimates for all the cells, we look for a statistical procedure which explore the dependencies between the variables in search for a parsimonious, yet reliable model. In addition, the use of latent variables have become increasingly common in this context, for example, in cases where there are several sources available that register the value of a variable of interest at an individual level, but none of them can be considered as error-free. In those cases, a latent categorical variable can be used to model the unobserved status. Since all involved variables are categorical we consider a family of log-linear latent class models which admit any structure of dependence between the variables. Estimates within this class can be extremely sensitive to model specification, particularly in presence of complex dependence structure and sparse data in small domains. To overcome the difficulty of model selection in such cases we propose a Bayesian model averaging approach to the problem. Usually a full Bayesian approach to model averaging requires the use of a Reversible Jump algorithm which is in general hard to implement. There are example of use of this algorithm within the class of log-linear models (without a latent variable). In this work, we show that, if we restrict ourselves to the subclass of decomposable models, it is possible to implement a simple Gibbs based MCMC. Some preliminary results seem to show that the restriction to that subclass does not affect the validity of the procedure.

IS15: Inference under informative sampling

Informative or ignorable selection process: a review

Daniel B. Bonnery
University of Cambridge, dbb31@cam.ac.uk

Isn't selection always informative and how to account for it ? In survey sampling, the observations are the outcome of two random processes: a process of interest and a selection process, called the nuisance process. The distribution of the second conditionally on the first is a function of a nuisance parameter. Informative selection occurs when selection should be accounted for when making inference. The different mathematical translations of this heuristic definition of informative selection found in the scientific literature will be reviewed in this presentation. The differences between the definitions and the gaps will be listed. A consensual general definition of the notions of ignorability, informativeness and "at random" will be proposed. Once agreed on a general definition of an informative selection, comes the question of what information to use when making inference. For example: is sample size informative ? is number of draws of the same unit in sampling with replacement informative ?. This will be the occasion to link this question to the fundamental and historical problem of sufficiency and ancillarity in presence of a nuisance parameter. In absence of a nuisance parameter, it is well known that the estimation of the model parameter should only be a function of a sufficient statistic. In presence of a nuisance parameter, there does not necessarily exist a Barndorff cut, and inference must be made on the nuisance process and the process of interest simultaneously. The presentation will provide a historic overview on how this problem was viewed, especially by R. Fisher, D. Basu, D Rubin and A. Scott. Without being controversial, it will be explained that the mathematical answer to practical questions about what information to discard or use, such as weighting or not weighting ? what elephant to weight and what estimator to choose in Basu's elephant fable ? should we borrow strength from all areas ? Is the selection informative ? are functions of the model chosen, so the real question is rather "what model should be used ?". Finally a review of the different techniques to make inference in presence of an informative selection process and illustrated with a practical application on crowd sourcing data.

Spatial processes and endogenous spatial selection, estimation and prediction

Francesco Pantalone, *University of Perugia, francesco.pantalone@studenti.unipg.it*
Daniel B. Bonnery, *University of Cambridge, dbb31@cam.ac.uk*
Maria Giovanna Ranalli, *University of Perugia, giovanna.ranalli@unipg.it*

This presentation applies the concept of informative selection, population distribution and sample distribution as described by Prof. Pfeffermann in a spatial process context. In this context, the output of the random process of interest does no longer consist of independent and identically distributed realisations for each unit of a population: the population is a continuous space and the observations are spatially dependent. A spatial selection process depending on the process of interest is considered. This selection induces a different dependence among selected units than the one in the population. We show how the sample distribution differs from the population distribution, and how one can account for this effect when doing statistical inference. The presentation focuses on Semi-variogram estimation, as well as prediction for non observed population units.

An Approximate Best Prediction Approach to Small Area Estimation for Sheet and Rill Erosion under Informative Sampling

Emily Berg
Iowa State University, emilyb@iastate.edu

The National Resources Inventory, a longitudinal survey of characteristics related to natural resources and agriculture on nonfederal US land, has increasingly received requests for sub-state estimates in recent years. We consider estimation of erosion in sub-domains of the Boone-Raccoon River Watershed. This region is of interest for its proximity to intensively cropped areas as well as important waterbodies. The NRI application requires a small area prediction approach that can handle nonlinear relationships and appropriately incorporate survey weights that may have nontrivial relationships to the response variable. Because of the informative design, the conditional distribution required to define a standard empirical Bayes predictor is unknown. We develop a prediction approach that utilizes the approximate distribution of survey weighted score equations arising from a specified two-level superpopulation model. We apply the method to construct estimates of mean erosion in small watersheds. We investigate the robustness of the procedure to an assumption of a constant dispersion parameter and validate the properties of the procedure through simulation.

Solicited Session 1

Small Area Estimation of Monetary Poverty in Mexico using Satellite Imagery and Machine Learning

David Newhouse, *World Bank Group, dnewhouse@worldbank.org*
Anusha Ramakrishnan, *World Bank Group, aramakrishnan@worldbank.org*
Tom Swartz, *Impira, tomgswartz@gmail.com*
Joshua D. Merfeld, *KDI School of Public Policy and Management, jdmesp@gmail.com*
Partha Lahiri, *University of Maryland, parthalahiri2@gmail.com*

This article investigates the accuracy and precision of small area poverty estimates for Mexican municipalities generated by combining survey data with auxiliary data derived from high-resolution satellite imagery. The auxiliary data are predictions of poverty rates and land classification categories generated by convolutional neural networks. Estimates are evaluated against official municipal poverty estimates derived from the 2015 inter-census survey. The preferred small area poverty estimates are obtained from a household-level empirical best predictor model of normalized per capita income linked with auxiliary data at the sub-area level. The resulting estimates are more accurate and far more precise than the direct survey estimates. For sampled municipalities, a household-level model yields more accurate and precise estimates than a sub-area model, which in turn produces more accurate and precise estimates than an area-level model. Estimates for sample municipalities are substantially more accurate and precise than those for out-of-sample municipalities. These results are robust to the use of an alternative sample from a different year. In simulations using area-level covariates, household model estimates are as accurate as area-model estimates for sampled municipalities and significantly more accurate for non-sampled municipalities. The household model estimates remain less accurate than official census-based estimates for 2010, especially for non-sampled municipalities. Combining survey and sub-area level satellite data using household-level empirical best models, while not always preferable to older census-based poverty estimates, significantly improves the accuracy and precision of survey-based estimates of monetary poverty.

Incidence of poverty in Costa Rica: small area estimates under a Structure Preserving Estimation (SPREE) approach

Alejandra Arias-Salazar
Freie Universitat Berlin, alejandra.arias@fu-berlin.de

Obtaining reliable estimates in small areas is a challenge because of the coverage and periodicity of the data collection. Several techniques of Small Area Estimation (SAE) have been proposed to produce quality measures in small domains by combining survey and census data at unit level under the assumption that both were conducted at (approximately) the same time, nevertheless, especially in developing countries censuses are usually carried out every ten years. Structure Preserving Estimation (SPREE) methods offer the possibility of producing intercensal updated cross-classified counts for small domain counts (Purcell and Kish (1980)). In this study, a new alternative is proposed by combining the properties of two of the most recent SPREE-type models: The Extended-SPREE (Isidro et al. (2016)) and the Multivariate-SPREE (Luna (2016)). The proposed method is used to obtain and update estimates of the incidence of poverty in Costa Rican cantons for six postcensal years (2012 - 2017). As uncertainty measure Mean Squared Errors are estimated via bootstrap and the adequacy of the proposal is also assessed by comparing with available regional-level poverty indicators.

Leave No One Behind: SDG Monitoring using Small Area Estimation in Latin America

Andres Gutierrez, *UNECLAC, andres.gutierrez@un.org*
Xavier Mancero, *UNECLAC, xavier.mancero@un.org*
Felipe Molina, *UNECLAC, felipe.molina.de@gmail.com*

The United Nations Economic Commission for Latin America and the Caribbean (UNECLAC) regularly analyses and presents household survey data. However, traditional data sources used to produce official statistics face several limitations to provide information for disaggregated population groups. To address the challenge of leave no one behind, required for implementing the Sustainable Development Goals (SDG) and the achievement of the 2030 Agenda, UNECLAC is working with countries statistical offices and other institutions to adopt SAE methods that combine information from different data sources (household surveys, population censuses, and administrative records) to produce geographically disaggregated official statistics. In this talk, we show some real applications of SAE methods for estimating disaggregated poverty indicators in some Latin American countries through a combination of household surveys and population censuses, following the EBLUP approach using both area-level and unit-level models. Based on these experiences, we summarize the challenges ahead to build capacities in Latin American countries regarding small area estimation and the Sustainable Development Goals monitoring.

Solicited Session 2

Skew-Normal CAR models for small domain estimation in the Brazilian Annual Service Sector Survey

Andre Felipe Azevedo Neves, *Brazilian Institute of Geography and Statistics, andrefelipe.neves@yahoo.com.br*
Denise Britz do Nascimento Silva, *National School of Statistical Sciences, denisebritz@gmail.com*
Fernando Antonio da Silva Moura, *Federal University of Rio de Janeiro, fmoura@im.ufrj.br*

In this paper we develop and evaluate skew-normal CAR models to produce state level gross service revenue estimates by economic activities for the North, Northeast and Midwest regions of the country. Since these domains have not been accounted for in the Brazilian Annual Service Sector Survey sampling plan, models are investigated to yield estimates with acceptable precision, particularly those that incorporate the dynamics over time and the similarity between economic activities. We propose normal asymmetric models to handle skewed economic data that bring together domain random effects, with time and state level random effects. In addition, a first-order random walk model and CAR models to borrow strength from related economic activities are considered. The results, based on 10-year survey data (2007-2016), show a substantial improvement in the precision estimates. Three asymmetric normal models merit attention when taking into account the trade-off between precision and bias of estimates besides model performance indicators: a model with domain random effects in the intercept and in the slope combined with a random walk effect; one with domain and time random effects in the intercept and in the slope; and a CAR model that incorporates activity structured random effects as well as domain and time random effects. Further investigation is underway to assess the use of model based estimates for the production of economic official statistics.

Estimation of the Employment Rate by Municipality in Mexico. Interpreting Results from an SAE model

Enrique de Alba
INEGI, enrique.dealba@inegi.org.mx
Eric Rodriguez
INEGI, eric.rodriguez@inegi.org.mx

In the first quarter of 2019 there were 2457 municipalities in Mexico, distributed among 32 states. The size of the total population in each one is very variable. The working age population (older than 15 years, according to Mexican legislation) in these municipalities goes from 68 people in Santa Magdalena Jicotitlán, Oaxaca to 1,448,788 in Iztapalapa, Mexico City. In this context, the analysis of results from applying small area estimation models to estimate unemployed population in each one of these municipalities requires identifying auxiliary variables that provide sensible results. In a study of the rate of employment by means of estimating both employed and total economically active populations by municipality, the resulting model shows an employment rate equal to, or larger than 99 percent of the economically active population for 328 municipalities, whose population older than 15 varies from 167 to 85,875 persons. Economic theory about the natural rate of unemployment indicates that in any economy there exists a subset of the total economically active population, which is part of frictional unemployment, defined as those who are moving from one job to another or that just entered the labor market. This makes it very unlikely to have occupation rates of 100 percent, where all people interested in getting a job have one, since this would motivate part of those population in the economically unactive who are available to work, who could try to enter in labor market. In this paper we look for variables that allow us to identify results from the model that are adequate to show the reality of the unemployment rate at the municipality level.

Small Area Estimates of Labor Force Statistics in Urban Mexico using Geospatial Data

Joshua D. Merfeld, *KDI School of Public Policy and Management, merfeld@kdis.ac.kr*
David Newhouse, *World Bank Group, dnewhouse@worldbank.org*
Micheal Weber, *World Bank Group, mweber1@worldbank.org*
Partha Lahiri, *University of Maryland, plahiri@umd.edu*

This article asks whether small area estimation with geospatial auxiliary data can successfully predict municipality-level labor force statistics in Mexico. We estimate labor force participation and unemployment separately for men and women using a synthetic survey derived from the 2020 census. We use two separate sampling strategies to create the survey: a simple random sample from randomly selected subareas as well as a full enumeration from those same subareas. After estimating a standard nested-error regression and modeling the outcomes as a linear function of geospatial data derived from publicly available sources we then evaluate the resulting estimates against the 2020 census data for both in-sample and out-of-sample municipalities. Incorporating the geospatial data substantially improves accuracy and precision. Both small-area estimates perform decidedly better than direct estimates for unemployment, which is quite low throughout the country and poorly estimated directly from the survey. However, we see the biggest gains when using a full enumeration of sampled subareas and in in-sample municipalities. For both labor force participation and unemployment, correlations with actual census values in these municipalities is above 0.8 and markedly better than direct estimates, at between 30 and 50 percent higher for labor force participation and is more than twice as high for unemployment.

Solicited Session 3

Flexible Small Area Estimation of Theil Index using Mixture of Beta

Silvia De Niccolò, *Università di Padova, silvia.denicolo@phd.unipd.it*
Maria Rosaria Ferrante, *Alma Mater Studiorum University of Bologna, maria.ferrante@unibo.it*
Silvia Pacei, *Alma Mater Studiorum University of Bologna, silvia.pacei@unibo.it*

This aim of the paper is to propose a small area estimation strategy for Theil Index, an entropy-based measure generally used to quantify economic inequality, industrial concentration and the disparity related to other economic phenomena, though it has also been used to evaluate racial segregation. We developed an area-level model of its relative index, i.e. Theil index over its maximum, which has a more manageable support between 0 and 1. Classical proposals in area-level context for measures on [0,1] are mostly based on proportions modelling and show limitations when dealing with asymmetric heavy-tailed data, complex variance functions and distributions close to the boundaries of the support, such as in our case. We propose a model with alternative distributional assumptions based on Simplex regression, estimated under a Hierarchical Bayes approach. An application to IT-SILC income data is provided, showing that our proposal yields a more flexible framework in comparison with Beta regression and classical gaussian area models with unmatched sampling and linking models, avoiding to underestimate income inequality. The model-based estimates show an average reduction of the coefficient of variation across areas around 25%, reaching in some cases the 56%.

On properties of MSE estimators of the EBLUP for some class of Linear Mixed Models in small area estimation

Malgorzata Krzciuk
University of Economics in Katowice, malgorzata.krzciuk@uekat.pl

The problem of small area prediction using the Empirical Best Linear Unbiased Predictor is considered. The proposed model assumes the correlation between vectors of random effects. Mean values in domains are predicted. The aim of the paper is the comparison of the properties of MSE estimators for the considered predictors. In the analyses we take into account i.e. classic estimator and estimators based on the parametric bootstrap method. The Monte Carlo simulation analyses based on a dataset from Local Data Bank (Statistics Poland) are prepared in R language.

Inference on quantiles in small area based on estimates of the distribution function

Tomasz Stachurski
University of Economics in Katowice, tomasz.stachurski@ue.katowice.pl

In economic studies, researchers are often interested in estimation measures of the position including median or other quantiles. In the case of positively skewed distributions, it is far more appropriate to use median rather than arithmetic mean as a measure of location. That is why quantiles have so many applications in the area of measuring poverty and income inequality. In small area estimation, it is quite common to use some auxiliary information in the process of estimation. Using such information can lead to better estimation precision, provided that auxiliary variables are highly correlated with the variable under study. There are a few approaches to estimate quantiles in survey sampling, including design-based and model-based approaches. The groundbreaking paper in the area of model-based quantile inference was Chambers and Dunstan (1986). This model is used to estimate the distribution function for non-sampled elements. Then, quantiles are achieved through inverting estimates of the distribution function. The essence of the problem of the estimation of the distribution function lays in obtaining predicted values of the study variable for non-sampled units assuming a linear mixed model. In the paper, we present a new proposition how to obtain predicted values for non-sampled elements in a small area. Based on the estimates of the distribution function, we get population and subpopulations quantiles as the inverses of the distribution function. Properties of proposed modifications are verified in a simulation study.

Solicited Session 4

Using Random Forests in SAE

Patrick Krennmair
Freie Universitat Berlin, patrick.krennmair@fu-berlin.de
Timo Schmid
Otto-Friedrich-Universitat Bamberg, timo.schmid@fu-berlin.de

Small Area Estimation (SAE) aims to construct reliable estimates for spatially dissaggregated indicators in the presence of small survey-sample sizes. This paper scrutinizes the use of random forests as versatile tools in the context of SAE for the estimation of area-level means. Model-based methods of SAE are predominantly conceptualized within the regression-setting and conventional methods rely on linear mixed models (LMM). Supplementary to linear and parametric models, machine learning methods offer non-linear and non-parametric alternatives, combining excellent predictive performance and a reduced risk of model-misspecification. This paper promotes the use of tree-based methods within the methodological tradition of SAE and identifies conceptual parallels. The direct application of random forests for SAE violates the independence assumption of errors. Mixed Effects Random Forests (MERF) (Hajjem et al., 2014) combine advantages of regression forests with the ability to model structural characteristics of data-dependencies. This paper provides a coherent framework easing the application of tree-based machine learning methods in SAE, including the development of a non-parametric mean squared error (MSE) bootstrap estimator. We compare the performance of proposed methods for point- and uncertainty-estimates using a model-based simulation study under complex scenarios. Additionally, we conduct a design-based simulation study, using real income data.

The comparison of different machine learning methods in small area prediction problems

Adam Chwila
University of Economics in Katowice, achwila@gmail.com

Nonparametric machine learning techniques have become very popular methods applied in many different areas of science. The different models such as: multivariate adaptive regression splines, gradient boosting regression trees, neural networks and support vector regression are the considered approaches, that are compared with standard small area approaches such as linear mixed models. The different models may produce different quality estimations in dependence on the data setup, which can be controlled by simulation settings. The predicted variable of interest is the disposable income of households. The disposable income for the most part of the society is strictly connected to the salaries, which are one of the most important working conditions measures. The subpopulations characteristics of disposable income such as mean, median, standard deviation, quartile deviation, moment skewness coefficient and quartile skewness coefficient are predicted. The performance of the considered models is compared also for the single households from the considered population. The simulation studies are based on the AMELIA universe dataset. Session

Statistical data integration as an extension of small area estimation for employee compensation

Andreea Luisa Erciulescu
WESTAT, AndreeaErciulescu@westat.com

Considered in this paper is the case where two surveys collect data on a common variable, with one survey being much smaller than the other. The smaller survey collects data on an additional variable of interest, related to the common variable collected in the two surveys, and out-of-scope with respect to the larger survey. Estimation of the two related variables is of interest at domains defined at a granular level. We propose a multilevel model for integrating data from the two surveys, by reconciling survey estimates available for the common variable, accounting for the relationship between the two variables, and expanding estimation for the other variable, for all the domains of interest. The model is specified as a hierarchical Bayes model for domain-level survey data and can be described as a macro-level data integration extension of a bivariate Fay-Herriot model. Posterior distributions are constructed for the two variables of interest. A synthetic estimation approach is considered as an alternative to the hierarchical modeling approach. The methodology is applied to wage and benefits estimation using data from the National Compensation Survey and the Occupational Employment Statistics Survey, available from the Bureau of Labor Statistics, Department of Labor, United States.

Solicited Session 5

Discovering Dynamics in Land Systems using Time Series Analysis and Non-linear Dynamical Methods

Richard Aspinall, *James Hutton Institute, rjaspinall10@gmail.com*
Michele Staiano, *University of Naples Federico II, mstaiano@unina.it*
Diane Pearson, *Massey University, D.Pearson@massey.ac.nz*
Alfonso Piscitelli, *University of Naples Federico II, alfonso.piscitelli@unina.it*

Dynamics in land systems are associated with process - response (cause and effect) relationships, endogenous behaviours resulting from system-level interactions, and adaptation to exogenous factors. Discovering these dynamics for different land uses, geographical contexts, and historical periods is a core activity of land systems science. In this paper, we use time series analysis, including methods from analysis of non-linear dynamical systems, to separate different types of dynamics encapsulated within a historical record of land system states for farming in Scotland over the period from 1867 to 2020. The data are from the annual series of agricultural census, that provides a summary of farming in Scotland. Although these data record the annual state of farming, they are seldom analysed for trends beyond short-term changes or for information beyond the status of different components of the funds of land use, livestock numbers and productivity that define the national account of farming. We use the data at a national scale, with a simple systems model of farming land use as a coupled human-environment system, to elucidate information on dynamics of farming and its evolution over time. Our results characterise dynamics from internal feedbacks and coupling of farming as a system at the national scale, reveal some system characteristics and behaviours associated with the dynamical evolution of farming as a system, and identify some regime shifts over the full 154-year timespan of the census. Specifically, the results reveal i) consequences of several exogenous factors as events that had an impact on system states, ii) show that arable and pastoral farming, at a national scale, are dynamically related over a range of timescales and coupled to global trends, and ii) that throughout much of the timespan of the study the system has maintained a pattern of changes consistent with endogenous systems-level feedbacks between sectors that act to dampen the impacts of exogenous factors. Changes in system dynamics over the timespan are also associated with policy changes that altered the interaction of arable and pastoral farming. Our analysis is based on the contention that the time series of system states recording the history of land use contain an embedded record of the impacts of long-, medium-, and short-term dynamics associated with both endogenous system forces and exogenous factors

that have influenced the land system. Because of this, both the underlying systems framework structuring the land system and the temporal scales at which a land system is studied should be made explicit, as the information needed for explanation of changes and dynamics will vary with the system structure and the time scales of interest. The use of time series analysis and methods from non-linear dynamics forces explicit attention to system structure, timescales and the multi-scale behaviours of land systems.

Small Area Estimation of Growing Stock Volume with Fay-Herriot area-level model

Aristeidis Georgakis
arisgeorg@for.auth.gr
Arne Nothdurft
University of Natural Resources and Life Sciences Vienna,
arne.nothdurft@boku.ac.at

The sustainability of the forests is supported by forest management plans. Respectively, forest management is based on management forest inventories (MFIs) that provide estimations and measurements of various forest parameters, with Growing Stock Volume (GSV) to be traditionally the most important. In recent years there is an increasing effort to produce more reliable estimates in management units (stands, compartments) than direct estimates which rely only on a few sample plots, utilizing auxiliary variables like data from the past or remote sensing data, and applying estimation techniques that make use of these like Small Area Estimation (SAE) statistical models. In this work, we explore the effectiveness of Fay-Herriot (FH) area-level models to produce small area statistics in the university forest of Pertouli for forest management purposes. The prediction variable dataset consists of sample plots, while census and remote sensing data serve as auxiliary explanatory variables in the estimation procedure.

Solicited Session 6

On benchmaking small area estimators when the model is misspecified

Renato Salvatore, *University of Cassino and Southern Lazio, rsalvatore@unicas.it*
Maria Chiara Pagliarella, *INAPP, mc.pagliarella@inapp.org*
Laura Marcis, *University of Cassino and Southern Lazio, laura.marcis@unicas.it*

In the context of small area estimation, benchmarking is justified by the need for adjusting individual area level estimates to agree with direct estimates of a larger area. The Eblup estimators do not satisfy the benchmarking property, and thus, in the last decades, many authors studied a variety of benchmarking tecniques in order to address this issue. In general, these methods rely on some modification of the Eblup by simple adjustments, as for the ratio and the difference benchmarking estimators (Steorts and Ghosh, 2013). Otherwise, an optimal benchmarking estimator that is model unbiased and at the same time satisfies the design-consistency property were obtained by Wang et al. (2008). Bell et al. (2013) give a general result for the optimal estimator in case of multiple benchmarking constraints, by joining together external and internal benchmarking using a common relation. Under model misspecification, Wang et al. also proposed an augmented model, by inserting a sampling variance model covariate, adjusted by the proportion of units in the corresponding area. Simulation experiments has shown that the augmented model estimator performs well in case of model misspecification, when the omitted variable is correlated with the augmented covariate. Nevertheless, self-benchmarking as in the You and Rao approach (2002) generally ensures efficiency in terms of the MSE, when direct estimates for the larger area suggest a model failure. By a general approach with multiple benchmarking constraints, this paper introduces a benchmarking linear estimator, assuming a model misspecification. The underlying assumption is that direct estimates for the larger area accounts for the true model. Then, we propose an augmented model that incorporates model failure. We show that misspecification is proportional to the orthogonal projection of the direct estimate in the subspace of the benchmarking constraints. Simulation experiments are given, in order to assess the validity of this approach.

Hierarchical Bayesian Spatial Small Area Model for Binary Data Under Spatial Misalignment

Kindie Fentahun Muchie, *Pan African University Institute for Basic Sciences, Technology and Innovation, muchie.kindie@students.jkuat.ac.ke*
Anthony Kibira Wanjoya, *Jomo Kenyatta University of Sciences, Technology and Innovation, awanjoya@gmail.com*
Samuel Musili Mwalili, *Jomo Kenyatta University of Sciences, Technology and Innovation, samuel.mwalili@gmail.com*

Small area model has become a popular method for producing reliable estimates for small areas. Small area modeling may be carried out via model assisted approaches within the design-based paradigm or model based approaches. A model assisted design based inference may be reliable in situations when there are large or medium samples in areas, while if data are sparse, model-based approach may be a necessity. Model based Bayesian analysis methods are becoming popular for their ability to combine information from several sources as well as taking account of uncertainties in the analysis and spatial prediction of spatial data. However, things become more complex when the geographic boundaries of interest are misaligned. Some authors have addressed the problem of misalignment under hierarchical Bayesian approach. In this study, we developed non-trivial extension of existing hierarchical Bayesian model for binary data under spatial misalignment. In this study, we developed a spatial hierarchical Bayesian small area model for a binary response variable under spatial misalignment. Fusion model, considering both areal level and unit level latent processes, was considered. The process models generated from the predictors were used to construct the basis so as to alleviate the well-known problem of collinearity between the true predictor variables and the spatial random process.

The inverse sampling method in the Big Data Era

Daniele Cuntrera, *Università di Palermo, d.cuntrera@email.it*
Vincenzo Falco, *Università di Palermo,*
vincenzo.falco01@community.unipa.it
Ornella Giambalvo, *Università di Palermo, ornella.giambalvo@unipa.it*

Thanks to the development of new IT tools, in the new millennium a large amount of data was seen, that is data sets containing an amount of information unthinkable to obtain in the last century. The amount of data that is created and stored globally is growing rapidly. This means that collecting key information from data of any nature becomes extremely important. This paradigm shift has led to the apparent uselessness of some classical statistical techniques (first of all the inference), which have been designed on the hypothesis of small samples. With the availability of a large amount of data, seems that the need to carry out a sample survey does not exist. Actually the so-called Big Data hide many pitfalls: often the relevant relationships and information are flattened, the large amount of data risks hiding or highlighting relationships where they exist or do not exist. In addition to this, these new huge amounts of data are increasingly forcing companies to adopt solutions that involve significant costs and not short times for the analysis of this data. A properly extracted sample seems to be more informative and accurate than a mass of population data or a part of them chosen without any probabilistic criterion. The goal of this work is to use the inverse sampling method for Big Data. Also, a practical application will be provided that takes into account simulated data or a real specific data set.

Solicited Session 7

Estimation of life expectancy in small areas using big data from the municipal registry

Stefano Cervellera, *Municipality of Taranto, s.cervellera@comune.taranto.it*
Carlo Cusatelli, *University of Bari Aldo Moro; carlo.cusatelli@uniba.it*
Massimiliano Giacalone, *University of Naples Federico II, massimiliano.giacalone@unina.it*

Given the increasing importance of knowing life expectancy, with less aggregation than what published by ISTAT at provincial level, the paper aims to estimate it for smaller municipal and sub-municipal areas, by means of big data from Registry of Municipal Census Office. The UN, on 25th September 2015, approved the 2030 Agenda, with 169 Sustainable Development Goals in 17 domains, placing in SDG 3 (Health and well-being) as many as 6 goals for mortality reduction out of 9 total and for improvement of life expectancy at birth and healthy life. In a particular moment for the reality of Taranto, squeezed between health, environment and justice, we highlight the validity of a tool, such as the real-time mortality observatory, which we have been designing for some years as a support to both medicine and criminal justice, as an alternative to the classic epidemiological analysis and appraisals which unfortunately are based on outdated information. In particular for the city of Taranto, where the environmental pressure is actually very strong, different statistical sources can be used to shorten these waiting times compared to ecological and epidemiological studies based on official data which have a longer circuit.

Reliable event rates for disease mapping

Harrison Quick
Drexel University, harryq@gmail.com
Guangzi Song
Drexel University, gs556@drexel.edu

When analyzing spatially referenced event data, the criteria for declaring rates as "reliable" is still a matter of dispute. What these varying criteria have in common, however, is that they are often not satisfied for crude estimates in most small area analysis settings, prompting the use of spatial statistical models to yield more reliable estimates. While this is a reasonable approach, recent work has quantified the extent to which popular models from the spatial statistics literature – namely the conditional autoregressive model – can overwhelm the information contained in the data, leading to oversmoothing and estimates that are likely far more precise than intended. In this study, we provide a definition for a "reliable" estimate for event rates that can be used for crude and model-based estimates and allows for discrete and continuous statements of reliability. By anchoring this definition in a Bayesian framework, we allow users to infuse prior information into their models to improve the reliability of their estimates while also providing guidance for how to control the informativeness of the model relative to the information contained in the data. After assessing the properties of our definition via a simulation study, we apply our approach to vital statistics data from the Commonwealth of Pennsylvania. Here, we highlight the effect of oversmoothing in spatial models with regards to our definition of reliability and how – when the model's informativeness is properly restricted – the definition of reliability proposed here can allow users to better focus their attention to areas where inferential decisions can be made with greater confidence. We then conclude with a brief discussion of how this definition of reliability can be used in the design of small area studies.

Solicited Session 8

Design-based small area estimation: an application to the DHS surveys

Ruilin Ren
ICF International, ruilin.ren@icf.com

Small area estimation (SAE) techniques have received increasing attention in the survey sampling domain due to the increasing request of sub-regional level data for policy making and planning. This is also the situation faced by the Demographic and Health Surveys (DHS) of The DHS Program, an international project funded by United States Agency for International Development (USAID). The DHS Program has collected, analyzed, and disseminated high quality data on population, health, HIV, malaria, and nutrition through about 400 surveys in over 90 countries since 1984. The DHS Program is facing more and more requests from the host countries to produce key DHS indicators at sub-regional/district level. Direct estimates, especially for the total fertility rate (TFR) and childhood mortality rates (CMR) which ask large sample size, is not feasible for DHS surveys because a very large total sample size can cause data quality concerns. Therefore, increasing sample size is not a proper answer to the request of sub-regional data. The only solution is to use small area estimation techniques to produce reliable estimation for sub-regions. The DHS surveys has a circle of five years. This research explores a new methodology on SAE, a design-based methodology, based on survey data collected in a single target survey or in a series of similar surveys conducted in recent years from the same country. The idea is to create a survey domain which is a nearest neighbor of the small area by pulling clusters close to the small area geographically, demographically, and timely, within one survey or from a series of similar surveys conducted in recent years. Because people living geographically close may have similar demographic characteristics, even they live in different administrative regions or sub-regions; secondly, most of the DHS indicators change slowly in time, especially for TFR and CMR, combine data from two or more similar surveys conducted within The DHS Program in the same country may enhance the power of analysis and therefor can produce reliable small area estimations. A survey domain created in this way is easier to analysis based on the survey design, variance estimation and confidence intervals are straight forward. On the other hand, data from DHS surveys are much more reliable and may also be timely close to the target survey than other data sources such as census data and administrative records. A calibration procedure can be applied to the small area estimations to ensure that they can be aggregated to regional level and match with the regional level estimation of the target survey, this

consistence property is important for small area estimations that many of the small area estimation techniques did not pay attention to.

Design-based composite estimation of small proportions in small domains

Andrius Ciginas
Vilnius University, andrius.ciginas@mif.vu.lt

Traditional direct estimation methods are not efficient for domains of a survey population with small sample sizes. To estimate proportions in the domains, we combine the direct estimators and the regression-synthetic estimators supported by domain-level auxiliary information. For the case of small true proportions, we introduce the design-based linear combination that is a robust alternative to the empirical best linear unbiased predictor (EBLUP) based on the Fay–Herriot model. We imitate the Lithuanian Labor Force Survey, where we estimate the proportions of the unemployed in municipalities. We show that the considered design-based compositions and estimators of their mean square errors are competitive for EBLUP and its accuracy estimation.

Challenges and lessons learned in using small area estimation for official statistics "how could we help?"

Haoyi Chen
Inter-Secretariat Working Group on Household Surveys, chen9@un.org
Yongyi Min
United Nations Statistics Division, min3@un.org

While small area estimation method has long been used by researchers in producing more disaggregated data, the use of such method for official statistics has been limited. In an effort to help countries in producing more disaggregated SDG data using this technique, the Inter-Secretariat Working Group on Household Surveys, in collaboration with the Inter-Agency and Expert Group on Sustainable Development Goal Indicators (IAEG-SDGs), is producing a Toolkit on Small Area Estimation (SAE) for SDGs. The Toolkit provides broad guidance on steps of using small area estimation including assessing user needs and input data availability, carrying out analysis and evaluation exercises and communicating the results with users and policymakers. Whenever available, case studies are presented for indicators under each of the 17 Sustainable Development Goals. Main audience of the Toolkit is statisticians from National Statistical Offices and other institutions within the National Statistical System that are interested in using SAE for the monitoring of the SDGs. In preparing for this Toolkit we have been carrying out discussions with a number of National Statistical Offices to understand the needs for small area estimation at country-level, the challenges they face and lessons learned in using such method for official data production. The paper will document what we have learned from our conversations with countries and discuss the way forward for the global statistical community and the researchers to assist countries in this area.